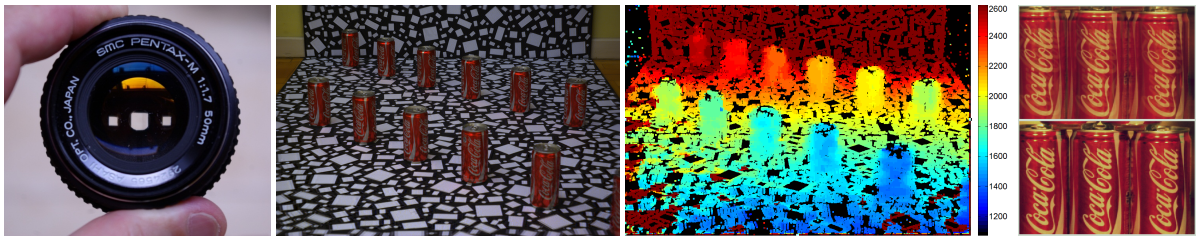
EUROGRAPHICS 2012 / P. Cignoni, T. Ertl
(Guest Editors)

Volume 31 (2012), Number 2

Coding Depth through Mask Structure

Horacio E. Fortunato† and Manuel M. Oliveira‡

Instituto de Informática – UFRGS



(a) (b) (c) (d)

Figure 1: *Our tri-lens implementation of a depth-encoding mask whose structural component consists of three Dirac deltas. (a) A tri-lens mask inside a Pentax 50 mm objective lens. (b) A structurally-deconvolved photograph of a scene captured through the lens shown in (a). (c) Distance map computed automatically from the captured image. The scale is in mm and illustrates the achieved accuracy and depth range. (d) Comparison between a portion of a captured image (top) and the corresponding structurally-deconvolved result (bottom).*

Abstract

We present a coded-aperture method based on a family of masks obtained as the convolution of one "hole" with a structural component consisting of an arrangement of Dirac delta functions. We call the arrangement of delta functions the structural component of the mask, and use it to efficiently encode scene distance information. We illustrate the potential of our approach by analyzing a family of masks defined by a circular hole component and a structural component consisting of a linear combination of three Dirac deltas. We show that the structural component transitions from well conditioned to ill conditioned as the relative weight of the central peak varies with respect to the lateral ones. For the well-conditioned structural components, deconvolution is efficiently performed by inverse filtering, allowing for fast estimation of scene depth. We demonstrate the effectiveness of our approach by constructing a mask for distance coding and using it to recover pairs of distance maps and structurally-deconvolved images from single photographs. For this application, we obtain significant speedup, and extended range and depth resolution compared to previous techniques.

Categories and Subject Descriptors (according to ACM CCS): I.4.1 [Computing Methodologies]: Digitization and Image Capture—Computational Photography.

1. Introduction

The use of optical masks to extend the limits of conventional photography has long been an active area of research. Introduced by Dicke [Dic68] and Ables [Abl68] in the context of X and Gamma-ray astronomy, the use of masks has become extremely popular in computational photography as a way to encode additional scene information. For instance, coded apertures have been used with single-lens systems to estimate scene depth [DC94, ZLN09], perform image deblurring [ZN09], motion deblurring [RAT06], and for per-

† e-mail: hefortunato@inf.ufrgs.br
‡ e-mail: oliveira@inf.ufrgs.br

forming deconvolution and estimating a coarse distance map from a single image [LFDF07].

The intuition behind the use of coded-apertures to encode depth is straightforward: an out-of-focus scene point p appears blurred on the camera's sensor; the blurring has the same shape as the camera's aperture, and *its size is proportional to the distance from p to the camera's focal plane*. Unfortunately, large amounts of blurring cause strong attenuations on high frequencies, which become difficult, if not impossible, to undo.

We present an efficient approach for recovering scene-depth information from single images based on a family of coded-aperture masks that can be modeled as the convolution of a set of delta functions and a single-hole mask (Figure 3). The arrangement of deltas functions, called the *structural component* of the mask, is used to encode distance information, allowing for the recovery of extended and more detailed distance maps when compared to traditional techniques. The *hole component*, used to improve the mask's light efficiency, is responsible for the blurring that arises from the finite size of the hole itself. Factoring the mask in these two components has several advantages, as it greatly simplifies mask analysis, design, implementation, and deconvolution. For instance, deconvolution can then be performed in two steps. First, the structural component is deconvolved producing a *structurally-deconvolved image* and a distance map. By properly designing the structural component (Section 4.2), such deconvolution can be efficiently performed using inverse filtering. This is much simpler than performing deconvolution with the original mask. The structurally-deconvolved image has the appearance of a regular photograph with a depth of field related to the size of the mask's hole component. A second (optional) step may be used to deconvolve the hole part, taking advantage of the distance map created in the first step.

We illustrate the potential of our approach by analyzing a family of masks whose structural components consist of a linear combination of three Dirac deltas disposed symmetrically around the origin. For such masks, we derive expressions for their structural-part inverses both in frequency and spatial domain. We show how such structural parts transition from well to ill conditioned as a result of varying a single parameter. We also discuss different ways of constructing actual masks with a given structural component. This contrasts with previous techniques where the masks are obtained performing an optimization on some searching space [RAT06, LFDF07, ZN09] and often require regularization even to recover the distance map.

We demonstrate the effectiveness of our approach by designing and physically constructing a mask for distance coding and using it to recover pairs of distance maps and structurally-deconvolved images from single photographs. Figure 1 (a) shows our mask prototype inserted in a Pentax 50 mm objective lens. Figures 1 (b) shows a structurally-deconvolved photograph from a scene captured using the modified lens in (a). The reconstructed distance map is shown in (c) with a scale in *mm*, illustrating a smoother and extended depth range compared to previous techniques [LFDF07]. Figure 1 (d) compares a portion of a captured image and the corresponding structurally-deconvolved result. Since our approach uses inverse filtering, constructing the depth map of a 10 Megapixel image takes 347 seconds using an unoptimized MATLAB script on a 3.0 GHz CPU.

The **contributions** of this paper include:

- An efficient approach for recovering scene-depth information from single images (Section 4). Compared to previous approaches, ours is significantly simpler, faster, and produces smoother and extended depth ranges;
- A family of ideal structural components consisting of a linear combination of three Dirac deltas disposed symmetrically around the origin. For these masks, we present formal derivations for their frequency and spatial-domain inverses, and an analysis of their noise amplification properties (Section 4.2);
- A demonstration of how to implement physical masks with a given structural-mask component (Section 5);
- A method for constructing masks using small lenses that decouples structural and hole components (Section 5.1).
- An algorithm for computing distance maps based on the analysis of the gradient magnitudes of structurally-deconvolved images (Section 5.1).

2. Related Work

Coded apertures were introduced by Dicke [Dic68] and Ables [Abl68] for imaging high-energy sources in astronomy. For this area, several mask patterns have been developed [Bro74, FC78, Fen78, Gol71, GF89], and a good survey can be found in [CSC*87].

Deblurring and Depth Coding: Coded apertures have recently received considerable attention in computational photography, where they have been used to improve image deblurring [RAT06, VRA*07, LFDF07, ZN09] and for coding scene-depth information [DC94, LFDF07]. Such techniques explore certain masks properties, such as their *high-frequency attenuation pattern* and the *presence of zeros in their Fourier transforms*. For instance, for image deblurring, Raskar et al. [RAT06] and Veeraraghavan et al. [VRA*07] search specifically for broadband masks, as these may simplify the deconvolution process; Dowski and Cathey [DC94], on the other hand, used coded aperture to obtain a distance map from a single image. Recently, Levin et al. [LFDF07] presented a technique that recovers a deblurred image and a distance map, also from a single image. Both approaches benefit from the use of masks with zeros in the frequency domain to improve distance discrimination. Bando et al. [BCN08] use color-filtered aperture to extract

depth and alpha matte, encoding depth as parallax differences among the color channels.

Most of these approaches use masks obtained through some optimization procedure over a regular grid of square holes with binary or varying transparency. Our approach, on the other hand, encodes depth using only the structural component of the mask. This leads to extended depth ranges, and allows for efficient structural deconvolution and depth recovery using inverse filtering. A second convolution step may be performed with the shape of the hole to improve deblurring.

Depth from Focus and from Defocus: These techniques extract depth from multiple images, and some variations use coded apertures to improve distance discrimination. Hiura and Matsuyama [HM98] present a multi-exposure camera and use simple pinhole delta patterns as coded masks. Zhou et al. [ZLN09] search for optimal coded-aperture pairs for depth from defocus. Levin [Lev10] analyses the problem of depth discrimination from a set of images captured with different coded apertures. The problem of finding optimal coded apertures for defocus deblurring is treated in [ZN09]. Contrary to these techniques, our approach extracts depth from a single image.

Light-field Capture: Several works explore light-field capture using arrays of micro lenses [Ng05, NLB*05, GZC*06, GI08], or a single multiplexed lens [LLW*08, NKZN08, NZW*10, GSMD07]. Ihrke et al. [IWH10] present a theory of plenoptic multiplexing. Several works [BEMKZ05, ERDC95, LHG*09] present techniques to extend the camera depth of field. We use small lenses to create physical masks with a given structural component. The images acquired with such masks, once structurally deconvolved have a depth of field given by the aperture of the small lenses.

3. Depth from a Single Image

The image of a scene point p located at a distance x in front of a thin lens is formed at a distance x' behind the lens (Figure 2). The relation between x and x' is given by Gaussian lens formula: $1/f = 1/x + 1/x'$, where f is the lens' focal length. If p is outside the camera's focal plane, its image appears blurred on the camera's sensor. The shape of the blurred image mimics the aperture mask, and its size depends on the point distance x. The diameters of the resulting circles of confusion on the sensor can be computed as:

$$d = L \left| s \left(\frac{1}{f} - \frac{1}{x} \right) - 1 \right|, \qquad (1)$$

where L is the diameter of the lens aperture, $|.|$ is the absolute-value operator, and s is the distance from the lens to the sensor (Figure 2). Equation (1) allows one to estimate the distance x from p's blurring size d. The process of image formation can then be modeled as

$$g = f \otimes m_x + \eta, \qquad (2)$$

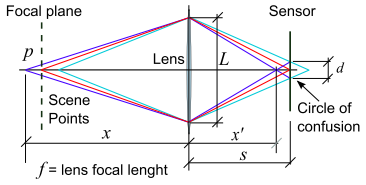


Figure 2: *The image of a point p at distance x in front of a lens is formed at a distance x' behind it. Points out of the focal plane appear blurred on the sensor. The size of the circle of confusion encodes the distance from p to the focal plane.*

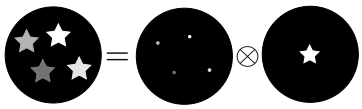


Figure 3: *Mask factorization: a multi-hole mask (left) expressed as the convolution of a structural component s_x (center) and a hole component h_x (right). s_x provides information about the holes' locations and transparency levels.*

where g is the captured image, and f is an ideal "all in focus" image. m_x is a mask that models the blurring of a scene point as a function of its distance x from the lens, and η models additive noise. Recovering scene depth from a single image can be achieved by selecting, for each image region, the mask scale that produces least artifacts in the deconvolved image [LFDF07, VRA*07].

4. Coding Depth through Mask Structure

Even though the whole mask may be used to code depth, we work with a family of masks that can be factorized into a structural component and a hole component, and show how to code distances using only the structural component. Masks factorable in this way are composed of several equally-shaped holes, with possibly different transparency levels. They can be expressed as the convolution of the *structural component* s_x and the *hole component* h_x (Figure 3):

$$m_x = s_x \otimes h_x.$$

The **structural component** specifies the locations and transparency levels of the holes. Thus, to identify the appropriate kernel size, only the structural component needs to be deconvolved. This has several advantages as s_x is easier to model and analyze than the whole mask m_x, and it can be designed for fast deconvolution. The **hole component** improves the mask's light efficiency. The structural component must not be confused with a physical pinhole mask. In our approach, s_x is always associated with a hole h_x.

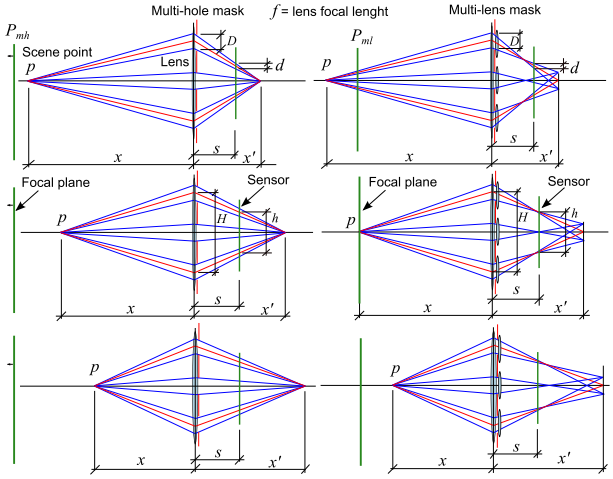


Figure 4: *Thin lens with a* multi-hole *(left) and a* multi-lens *mask (right). Structural component analysis (both designs): chief rays (solid red) from a scene point p at distance x from the thin lens converge on a plane at distance x' behind it. H (middle row) is the distance between the centers of two mask holes. h is the distance between the corresponding chief rays over the sensor. One can infer x from h using structural information only. Hole component analysis: D (top row) is the size of the mask holes. d is the size of the circle of confusion associated with a point p at distance x. Rays passing through the small lenses (right column) converge a little before than in the case of the multi-hole mask. The multi-lens design leads to smaller circles of confusion, which translates into extended depth range discrimination. It allows for the determination of the distances of scene points located behind, on, or in front of the camera's focal plane P_{ml}.*

4.1. Masks with a Given Structural Component

There are several ways of constructing real masks with a given structural component. Essentially, one can use any method that superposes images captured with the same hole-component mask slightly displaced and with controllable relative intensity. For instance, they can be implemented using a mask with several holes of the same shape and size, pinholes, small lenses, prisms, tilted parallel-sided glasses, combinations of multiple exposures, etc. This section describes two realizations: a *multi-hole* and a *multi-lens* mask.

A **multi-hole mask** consists of several equal holes cut out from a support. Figure 4 (left column) illustrates such a mask attached to a thin lens with focal distance f. The image of a point located at a distance x in front of the lens will be sharply formed at a distance x' behind the lens. If the camera sensor is at distance $s \neq x'$ behind the lens, multiple blurred images of the point will appear on the sensor. The relationship between x and the distance h between the centers of the

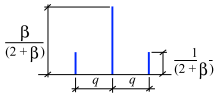


Figure 5: *The trident structural component. The central delta is β times bigger than the laterals ones. $1/(2+\beta)$ is a normalization factor. q is the distance between the lateral deltas and the central one.*

projected circles of confusion is given by

$$h = H \left| s\left(\frac{1}{f} - \frac{1}{x}\right) - 1 \right|. \tag{3}$$

Here, H is the distance between the centers of two holes in the mask (Figure 4, second row, left). The centers of all holes define the structural component of the mask. A similar relation can be derived for the sizes d of the circles of confusion:

$$d = D \left| s\left(\frac{1}{f} - \frac{1}{x}\right) - 1 \right|. \tag{4}$$

The **multi-lens mask** is shown in Figure 4 (right column). It has small convergent lenses on top of the mask holes. The focal length of the small lenses is much larger than the camera's main lens focal length. The small lenses do not deviate rays passing through their centers (chief rays), and Equation (3) for the structural component still holds (Figure 4, middle row). Due to the extra power of the small lenses, rays coming through an individual hole now converge a little before than in the case of the multi-hole mask. The diameters of the resulting circles of confusion on the sensor can be computed as:

$$d = D \left| s\left(\left(\frac{1}{f} + \frac{1}{f_{small}}\right) - \frac{1}{x}\right) - 1 \right|, \tag{5}$$

where f_{small} is the focal length of the small lenses.

In both designs, the distance x of a scene point p can be estimated from the dimension h of the projection of the structural component of the mask on the sensor (Figure 4).

4.2. The Trident: a Simple Structural Component

To make the discussion more concrete, we introduce a family of structural components consisting of three Dirac deltas disposed symmetrically around the origin. The weight of the central peak can be different from the laterals ones. It means that, when convolved with a hole, the "transparency level" of the central hole of the resulting mask will be different from that of the lateral holes. We call these structural components **tridents**. They are characterized by two parameters: the distance q from one lateral delta to the central one, and the relative magnitude β of the central delta with respect to the two lateral ones (Figure 5). β plays an important role in determining the invertibility of the resulting kernel.

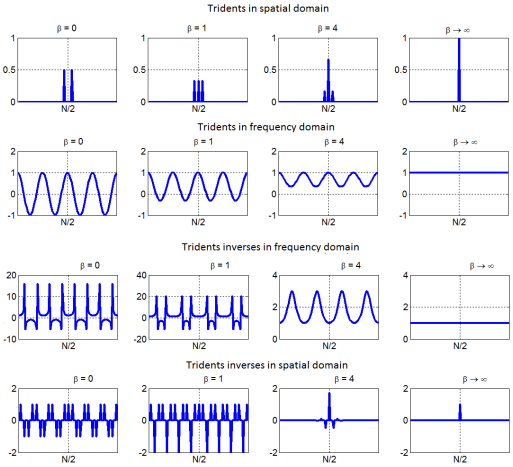


Figure 6: *Comparison of trident properties for various values of* β *(from left to right,* $\beta = 0$*, 1, 4, and* ∞*). For* $\beta > 2$*, the tridents have no zeros in the frequency domain. From top to bottom: spatial domain, frequency domain, inverses in frequency domain, and inverses in spatial domain.*

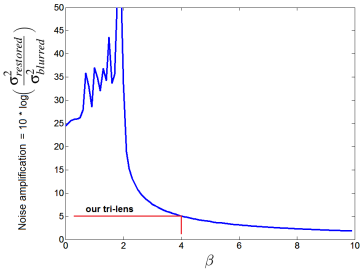


Figure 7: *Trident's inverse noise amplification* $(10\ log(\sigma^2_{restored}/\sigma^2_{blurred}))$ *as a function of* β *for a 500-column mask with* $q = 1$*. Our tri-lens mask (Section 7) uses* $\beta = 4$*.*

In spatial domain, a trident is represented by a matrix with only three non-null elements on the first row:

$$t^{(\beta,q)}_{0,0} = \frac{\beta}{(2+\beta)}, \quad t^{(\beta,q)}_{0,q} = t^{(\beta,q)}_{0,C-q} = \frac{1}{(2+\beta)}, \qquad (6)$$

where $t^{(\beta,q)}_{r,c}$ is the matrix element at row r and column c. C is the number of columns of the matrix. The Fourier transform of a trident is

$$T^{(\beta,q)}_{u,v} = \frac{\beta + 2\cos(2\pi(\frac{vq}{C}))}{(2+\beta)}. \qquad (7)$$

Appendix A presents derivations for this expression, for the trident inverse in frequency and spatial domains, and for the trident inverse noise amplification. According to Equation (7), for $\beta > 2$ the resulting trident has no zeros in the frequency domain. For such well-conditioned structural components, deconvolution can be performed using inverse filtering, which is simple and fast. The value $\beta = 2$ defines a transition from an ill-conditioned to a well-conditioned kernel (Figure 7). If $\beta = 0$, it reduces to a *two-delta kernel*. Likewise, as β approaches infinity, it converges to a single delta at the origin, which is the *identity operator*. Figure 6 shows a comparison of tridents for various values of β. Tridents with $\beta > 2$ have no zeros in the frequency domain. The transition between well-conditioned to ill-conditioned is clearly visible in Figure 7, which shows the trident's inverse noise amplification as a function of β for a 500-column mask and $q = 1$. A detailed explanation of noise amplification and the derivation of the plotted data is presented in Appendix A.

(a) (b) (c)

Figure 8: *A tri-lens mask. (a) one prototype built with 5.5 mm diameter contact lenses. (b) Mask installed inside a Pentax 50 mm objective lens. (c) View from behind the lens.*

5. Practical Implementation: The Tri-lens Mask

We constructed a multi-lens mask (Figure 4, right column) that has a trident structural component. We call this mask a **tri-lens mask** (Figure 8, (a)). We built it using three fluorocarbon custom-made contact lenses with focal distance of 4 meters (*i.e.*, +0.25 diopters) and 5.5 mm of diameter. They were mounted on a plastic support, with their optical axes parallel to one another and 6 mm apart. Behind the lenses, we added a black paper mask with three square holes aligned with the holes of the support. The central hole has 4x4 mm^2, while the two lateral holes have 2x2 mm^2, making $\beta = 4$. The use of a central hole larger than the lateral ones simplifies construction by reducing the number of components. The limitations of this implementation choice are discussed in subsection 7.4.

The choice of $\beta = 4$ guarantees low noise amplification (Figure 7), allowing deconvolution of the structural component to be performed using inverse filtering. There is a compromise between noise amplification (ease of inversion) and depth discrimination: larger β values reduce noise amplification, but attenuate the relative contribution of the lateral lenses, which reduces depth discrimination. For the experiments reported in this paper, we inserted the tri-lens mask on a 50 mm f:1.7 lens (model smc PENTAX-M 1:1.7), as shown in Figures 8 (b) and (c), which was attached to a 10 Megapixel PENTAX K10 DSLR camera. The mask was mounted inside the lens close to the diaphragm, which continues to operate freely.

5.1. Depth Estimation with a Tri-lens Mask

An image captured through a tri-lens mask consists of three slightly displaced copies of the scene. This is illustrated in the zoomed-in portion of the calibration panel shown in Figure 10 (bottom left). The displacement between the leftmost and the rightmost copies varies with the distance from the scene object to the camera and, within the thin lens approximation, is given by Equation (3). An image deconvolved with the wrong trident scale has artifacts not present in the image deconvolved with the correct trident scale (Figure 9). The sum of the magnitudes of the first derivatives computed at a neighborhood N_i around a pixel p_i in the deconvolved image has a local minimum at the correct trident scale. To associate a depth with a pixel p_i, we find the trident size that minimizes this sum at a neighborhood N_i around p_i, and use the calibration data of Figure 10 to obtain the corresponding depth. This corresponds to searching for the solution that minimizes an L1 norm on the gradients in the reconstruction. It can be interpreted as a sparse gradient image-prior assumption, similar to the one used in [WF07, LFDF07].

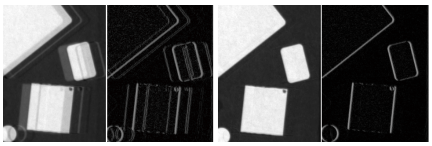
6. Extended Depth Range and Resolution

Depth estimation is based on blur-kernel size estimation. Thus, depth discrimination range and resolution are limited by the range of kernel sizes, measured in pixels over the sensor, that can be distinguished. Levin et.al. [LFDF07] report a range of discriminable kernel sizes from 4 up to ≈ 14 pixels. When the blur size is smaller than 4 pixels, depth discrimination becomes impossible due to the lack of structure; for blur sizes larger than 14 pixels, the blur cannot be robustly inverted. Veeraraghavan et.al. [VRA*07] report good deblurring results up to a blur size of ≈ 20 pixels. In contrast, our method handles much larger kernel sizes, which can vary from 25 to 75 pixels (Figure 10).

The reason for our extended depth-discrimination capability is twofold: first, the well-conditioned structural component is invertible for almost all scales (*i.e.*, the frequency attenuation properties of its Fourier transform only depend on the parameter β, not on the parameter q that defines the scale). The limit on the resolution range is imposed by the size of the hole component, which attenuates high frequencies, destroying information about image features. Second, our tri-lens design minimizes the size of the projection of the hole component in the central part of the depth discrimination range, where it vanishes. Besides generating smaller circles of confusion, it discriminates scene points that are in front, on, or behind of the camera's focal plane. There is no ambiguity about equidistant points at different sides of the focal plane.

7. Results

Figures 1, 11 and 12 show images processed using our approach. The images were captured using raw DNG (Digital Negative) format, which were then converted to ppm and processed using MATLAB scripts. The camera was mounted on a tripod and delayed shooting was used to minimize camera shake during capture. The scenes were indirectly illuminated with two 500 watt halogen light reflectors. The camera sensitivity was set to ISO 100, and shutter speed to 0.4 secs to compensate for the low intensity provided by indirect illumination. The camera aperture was fixed at f/1.7 to maintain the diaphragm completely opened and to not interfere with the tri-lens mask. We set the focal distance at 2.4 meters according to the graduated scale of the camera lens. The presence of the tri-lens mask changes the resulting focal distance. This setup corresponds to the upper curve in Figure 10, which gives a discernible range of distances from 1,400 to 2,600 mm.



(a) (b) (c) (d)

Figure 9: *Image deconvolved: (a) with wrong trident scale, and (c) with correct trident scale. (b) and (d) are the respective horizontal derivative magnitudes (contrast enhanced).*

7.1. System Calibration

The effective size of the tri-lens kernel on the sensor varies with object distance. However, discrepancies between the thin lens model and the real optical system also cause the kernel to vary its form across the image. Thus, we model the trident kernel using three parameters recovered through a calibration process:

- β: the relative intensity of the central delta with respect to the laterals ones;
- q_l and q_r: the distances from the left and right deltas, respectively, to the central one.

We use a semi-automatic calibration procedure based on the minimization of the sum of the magnitudes of the first derivatives (Section 5.1) of a deconvolved calibration pattern placed at known distances. The calibration pattern consists of a $1,000 \times 1,600\ mm^2$ black panel covered with randomly distributed white rectangles and circles (Figure 10 left). We then took pictures of the panel from distances varying from 830 mm to 3,000 mm at increments of 70 mm. The images were divided in 35 blocks (5 by 7), as shown in the bottom-left portion of Figure 10 (right). Each block was processed independently to obtain the optimal β, q_l and q_r parameters for the deconvolution kernel. The parameter values obtained for each image block and distance to the calibration pattern were linearly interpolated to provide a complete set of trident mask parameters for every pixel of an image at a given

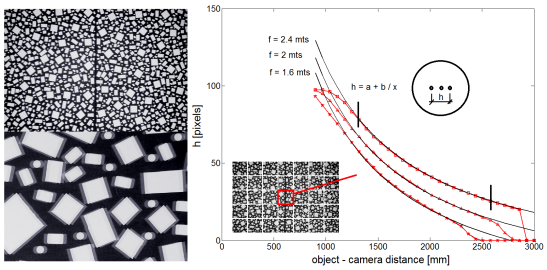


Figure 10: *(right) Trident size versus objects distances. Calibration curves for the green channel in the central region of the image for three focal distances. A calibration panel (left-top) and a zoomed-in portion of it as imaged through a tri-lens mask (left-bottom). This image consists of three slightly displaced copies of the pattern.*

distance in the calibration range. Since each color channel presents slightly different calibration results, they were calibrated independently. Thus, deblurring an RGB image requires three passes, one for each color channel.

Figure 10 (right) shows calibration curves obtained for three different focal distances according to the marks of the Pentax lens. It plots the distances between the lateral peaks (*i.e.*, $q_l + q_r$) for the green channel, inside the calibration range. The plotted data correspond to distances computed for the central block of the images. The solid lines represent fittings to the experimental data using an $h = a + b/x$ approximation, as suggested by Equation (3). Here, x is the distance from the calibration panel to the camera.

7.2. Distance Map Computation

Given an image I captured using a tri-lens mask, the process of recovering a distance map can be divided in four steps: (i) perform deconvolution of I with kernels of various scales s_i, producing a sequence of structurally-deconvolved images I_i; (ii) for each deconvolved image I_i, compute the sum D_{ij} of the absolute values of the first derivatives of I_i at a neighborhood N_j around each pixel p_j; (iii) identify the scale s_k that minimizes the D_{kj} for each pixel p_j; and (iv) discard unreliable distance values.

Deconvolution for depth map calculation is performed at full resolution (10 Megapixels) for 160 depth layers (increments of 8.75 mm in a range from 1,275 to 2,675 mm). The images are divided in blocks of 128 columns by 64 rows. Each block is deconvolved using one-dimensional inverse filtering in frequency domain. To avoid artifacts, deconvolution is done on an window with 512 pixels, after which only the central 128 columns are kept.

The second step consists of calculating D_{ij} at a neighborhood N_j around each pixel p_j. In practice, however, we only calculate D_{ij} at every 16 pixels both in the horizontal and vertical directions, thus creating a distance map with a resolution of $1/256$ of the resolution of the input image I. N_j consists of 80 rows by 144 columns of the deconvolved images I_i (at full resolution). For each pixel p'_k in the distance map, we assign the minimum D_{ik} (*i.e.*, the minimum D value for that pixel across all scales). As mentioned in the previous section, different color channels require independent calibrations. To compute distances maps we only process the green color channel, as that is the channel of best resolution.

As with any other passive depth-extraction technique, image regions with uniform luminance levels (*i.e.*, no texture) do not provide enough information for distance recovery. Thus, we discard distances corresponding to featureless regions of the image. For this, we built a binary mask based on the magnitude of the derivative at each pixel. To compute these derivatives, we use a luminance version of the captured image and a 1D horizontal (three-pixel-wide) high-pass filter. We further remove noisy pixels using morphological opening with a vertical-line structuring element consisting of 4 pixels. Then, we fill small gaps using morphological closing with a 20×20 pixels structuring element.

Processing time varies linearly with the number of distances (scales) processed. For example, for 160 depth layers, the algorithm takes a total of 347 seconds for a 10 Megapixel image using an unoptimized MATLAB script on a 3 GHz Core 2 Extreme CPU. This is significantly faster than the approach of Levin et al. [LFDF07], for which the authors report one hour to process a 2 Megapixel image on a 2.4 GHz CPU, for a only a few depth layers.

Figures 1, 11, and 12 show distances maps obtained for three different scenarios designed to test the limits of our system. Black marked regions correspond to unclassified pixels, which result primarily from the lack of texture in those regions. Figure 1 (b) shows an ordered array of cans positioned against two calibration panels. The panels were used to check the ability of our solution to produce smooth distance maps. The range of recovered distances is shown in Figure 1 (c) and smoothly varies from 1,200 mm to 2,600 mm, with the depth of the great majority of pixels correctly classified. Such a range is about twice as large as the one reported by Levin et al. [LFDF07] and much more detailed.

Figure 11 shows the same arrangement of cans placed against two sheets of textureless clear paper. For this example, our solution is capable of estimating depth values for the cans and for a seam between the two paper sheets, as well as for a visible edge on the right. As expected, most of the pixels were discarded due to the lack of texture information.

Figure 12 tests our solution with a natural scene with several objects placed within the range of discriminable distances. Our solution recovers proper distance values all along its calibration range for regions containing discernible features. Note its ability to exploit even subtle details, such as in the case of the dark statues in the back of the scene.

7.3. Structural Image Deconvolution

We use the distance map, as well as the calibration data to select the appropriate kernel parameters for structural image deblurring (deconvolution of the structural component). Since each cell of the distance map corresponds to a block of 16×16 pixels in the captured image I, all pixels in a block are deconvolved using the same kernel parameters. Deconvolution is performed by 1D inverse filtering in the frequency domain. To avoid artifacts, it is done on a 140-pixel-wide window but only the central 16 columns are kept. For featureless regions of the image, we simply copy the pixel values from the input image. Because of the separate calibration of the different color channels, they are processed independently. As we only deconvolve the structural part of the mask, deconvolution removes the three ghost-like effect introduced by the tri-lens mask. With our current prototype, structurally-deconvolved images look like conventional photographs taken with a small aperture (the size of the small lens) of approximately f/8.

Figures 1 (d) and 13 compare portions of captured images with the structurally-deconvolved results. Figure 13 shows a zoomed-in version of the structurally-deconvolved result obtained for a portion of the living-room scene shown in Figure 12. All features of these objects have been sharply reconstructed. For instance, compare the phone antenna and the figurines in the captured and deconvolved images. High-resolution versions of these images are provided in the supplementary materials for close inspection.

7.4. Discussion and Limitations

Our tri-lens mask prototype is a good approximation to an ideal mask that can be factorized into structural and hole components. It provides important insights about the behavior and potential of more complex masks. For instance, it shows that it is possible to discriminate distances in front, on, and behind the camera's focal plane. By orienting the trident so that it becomes horizontally (or vertically) aligned with the camera sensor, one simplifies the implementation of the algorithms required to process the captured data. Such algorithms can then be implemented in 1D along rows (columns) of the captured image. On the other hand, such a configuration does not allow the exploration of image gradients perpendicular to the direction of the trident. To take advantage of gradients along arbitrary directions, the mask can be modified so that its peaks form a triangle.

Creating a tri-lens mask requires aligning the small lenses on its support, which is not required for other masks used in computational photography [RAT06, VRA*07, LFDF07, ZLN09, ZN09]. The use of a central hole bigger than the lateral ones (used to get $\beta = 4$) causes the prototype to not exactly follow the image formation model, and may lead to the occurrence of ghost artifacts. However, such artifacts only appear for objects far from the focal plane of the optical system comprised by the camera lens and the small

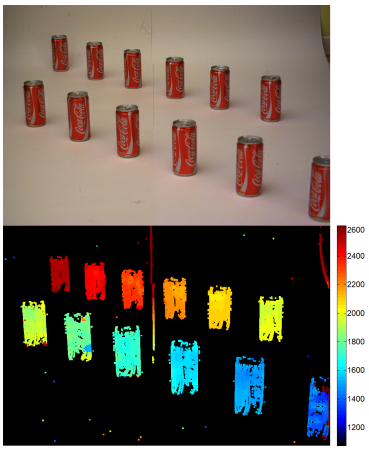


Figure 11: *Coke cans against a textureless background. (top) Structurally-deconvolved image. (bottom) Recovered distance map.*

lenses. This results from the superposition of "circles" of confusion of different scales on the camera's sensor, which cannot be undone through deconvolution by the inverse of a pure trident. For objects close to the focal plane, the differences in the scales of the circles of confusion can be disregarded, as demonstrated by the high-quality deconvolution result shown in Figure 13. One way of avoiding the occurrence of such artifacts is to use holes of the same size under the three small lenses, and enforce $\beta = 4$ using neutral density filters over the lateral holes. Artifacts may also result from errors in the calibration process in areas of the image away from its center. This is likely to be the main cause of the errors in the depth estimation at bottom left portion (bottle and chair) of Figure 12. Finally, modifying the camera lens' focal distance requires recalibrating the system for this new configuration, which is also required for conventional masks.

8. Conclusion and Future Work

We have presented an approach for the analysis and design of a family of coded-aperture masks for computational photography. Such masks can be modeled as a convolution between a structural component and a hole component. Factoring a mask into these elementary components greatly simplifies both analysis, design, implementation, and deconvolution. This results from the fact that some properties can be identified with individual mask components, and their analysis carried out separately from other aspects of the mask. We have analyzed a family of structural masks, for which we presented a formal treatment for their frequency and spatial domain inverses and corresponding noise amplification properties. We also discussed different ways of constructing physical masks with a given structural component. This con-

Figure 12: *Detail of a living room. (top) Structurally-deconvolved image. (bottom) Recovered distance map.*

Figure 13: *Portion of the living room scene from Figure 12. (left) Image captured with our tri-lens mask. (right) Structurally-deconvolved image. Note the properly reconstructed details on the antenna, phone base, and figurines.*

trasts with previous approaches, which are based on the use of optimization procedures [RAT06, LFDF07, ZN09].

We demonstrated the effectiveness of our approach by designing and physically constructing a mask for distance coding and using it to recover pairs of depth maps and structurally-deconvolved images from single photographs. Since our approach uses inverse filtering, it is significantly faster than previous techniques. Our structural-mask-based solution also allows for the recovery of smooth distance maps with extended range in comparison with previous approaches. For this, we use an algorithm based on the analysis of the gradient of the deconvolved images.

We have demonstrated how structural components can be used for encoding depth. There should be other problems that can be efficiently solved with the use of structural mask components. We hope that our approach will inspire researchers to design masks with optimal structural and hole components for solving other problems, as well as to analyze the properties of existing masks using the presented factorization technique.

Appendix A: Trident Analysis

In spatial domain, the trident is represented by a matrix with R rows and C columns and only three non-null elements on the first row (Equaton 6). Using the definition of the Fourier transform F of a matrix f with R rows and C columns:

$$F_{u,v} = \sum_{r=0}^{R-1}\sum_{c=0}^{C-1} f_{r,c}\exp\left(-2\pi i\left(\frac{ur}{R}+\frac{vc}{C}\right)\right),$$

The Fourier transform of the trident $t^{(\beta,q)}$ (Equation (6)) is:

$$T_{u,v}^{(\beta,q)} = \frac{1}{(2+\beta)}\left[\beta+\exp\left(-2\pi i\left(\frac{vq}{C}\right)\right)+\exp\left(-2\pi i\left(\frac{v(C-q)}{C}\right)\right)\right]$$

$$T_{u,v}^{(\beta,q)} = \frac{\beta+2\cos(2\pi\left(\frac{vq}{C}\right))}{(2+\beta)}. \qquad (8)$$

Its inverse in the frequency domain is simply the multiplicative inverse of Equation (8): $T_{u,v}^{-1(\beta,q)} = 1/T_{u,v}^{(\beta,q)}$.

We present formal expressions for the trident inverse $t^{-1(\beta,q)}$ in spatial domain for two representative cases of Equation (6): (i) for $q = 1$ (Equation 9) and (ii) for $q' > 1$ and C an integer multiple of q (*i.e.*, $C = nq$) (Equation 10). Due to space constraints, the derivations of these expressions are provided as part of the supplementary materials.

$$t_{i,j}^{-1\ (\beta,1)} = \delta_{i0}\ (-1)^j\frac{2+\beta}{exp(\lambda)-exp(-\lambda)}\exp\left(-\lambda j\right) \qquad (9)$$

$$t_{i,j'}^{-1(\beta,q')} = \begin{cases} t_{i,j}^{-1\ (\beta,1)} & \text{if } j' = q'\cdot j \\ 0 & \text{otherwise} \end{cases} \qquad (10)$$

Trident Inverse Noise Amplification: The pixel values of a noisy image can be seen as a set of independent random variables. Thus, the deconvolution of a blurred image B using a trident's inverse mask $t^{-1(\beta,q)}$ can be represented as a linear combination of these random variables. The variance $\sigma^2_{restored}$ in the pixel values of the deconvolved (restored) image can then be expressed as:

$$\sigma^2_{restored} = \sigma^2_{blurred}\sum_{r=0}^{R-1}\sum_{c=0}^{C-1}|t_{r,c}^{-1(\beta,q)}|^2, \qquad (11)$$

where $\sigma^2_{blurred}$ is the variance associated with the pixel values of the blurred image. Using Parseval's theorem, the same result may be obtained in the frequency domain:

$$\sigma^2_{restored} = \sigma^2_{blurred}\frac{1}{RC}\sum_{u=0}^{R-1}\sum_{v=0}^{C-1}|T_{u,v}^{-1(\beta,q)}|^2. \qquad (12)$$

When $\beta > 2$ and $C \to \infty$, it follows from Equations (9) and (11) that the resulting mean-square noise amplification is:

$$\frac{\sigma^2_{restored}}{\sigma^2_{blurred}} = \left(\frac{2+\beta}{exp(\lambda) - exp(-\lambda)}\right)^2 \frac{1}{1-\exp(-2\lambda)}. \quad (13)$$

According to Equation (11), noise amplification will be high when the inverse contains terms with high magnitude spread all over its domain, as shown in the two leftmost examples in the bottom row of Figure 6. In the frequency domain, this corresponds to the occurrence of zeros in the Fourier transform of the mask itself. For values of $\beta > 2$, the Fourier transforms of both the trident masks and their inverses contain no zero.

Acknowledgments

Horacio E. Fortunato was supported by a postdoctoral fellowship (Bolsa REUNI, PROPG-UFRGS). Manuel M. Oliveira acknowledges a CNPq-Brazil fellowship (No. 308936/2010-8). We wish to thank Optolentes (http://www.optolentes.com.br) for kindly providing the contact lenses used in our tri-lens prototype.

References

[Abl68] ABLES J. G.: Fourier transform photography: A new method for x-ray astronomy. vol. 1, p. 172. 1, 2

[BCN08] BANDO Y., CHEN B.-Y., NISHITA T.: Extracting depth and matte using a color-filtered aperture. *ACM Trans. Graph. 27* (December 2008), 134:1–134:9. 2

[BEMKZ05] BEN-ELIEZER E., MAROM E., KONFORTI N., ZALEVSKY Z.: Experimental realization of an imaging system with an extended depth of field. *Appl. Opt. 44*, 14 (May 2005), 2792–2798. 3

[Bro74] BROWN C.: Multiplex imaging with multiple pinhole cameras. 1806–1811. 2

[CSC*87] CAROLI E., STEPHEN J. B., COCCO G., NATALUCCI L., SPIZZICHINO A.: Coded aperture imaging in X- and gamma-ray astronomy. *Space Science Reviews 45*, 3 (1987), 349–403. 2

[DC94] DOWSKI E. R., CATHEY W. T.: Single-lens single-image incoherent passive-ranging systems. *Appl. Opt. 33*, 29 (Oct 1994), 6762–6773. 1, 2

[Dic68] DICKE R. H.: Scatter-hole cameras for x-rays and gamma rays. *Astrophysical Journal 153* (1968), L101. 1, 2

[ERDC95] EDWARD R. DOWSKI J., CATHEY W. T.: Extended depth of field through wave-front coding. *Appl. Opt. 34*, 11 (Apr 1995), 1859–1866. 3

[FC78] FENIMORE E. E., CANNON T. M.: Coded aperture imaging with uniformly redundant arrays. *Appl. Opt. 17*, 3 (Feb 1978), 337–347. 2

[Fen78] FENIMORE E. E.: Coded aperture imaging: Predicted performance of uniformly redundant arrays. *Applied Optics 17* (1978), 3562–3570. 2

[GF89] GOTTESMAN S. R., FENIMORE E. E.: New family of binary arrays for coded aperture imaging. *Applied Optics 28* (1989), 4344–4352. 2

[GI08] GEORGEIV T., INTWALA C.: Light field camera design for integral view photography, 2008. 3

[Gol71] GOLAY M. J. E.: Point arrays having compact, nonredundant autocorrelations. *J. Opt. Soc. Am. 61*, 2 (Feb 1971), 272–273. 2

[GSMD07] GREEN P., SUN W., MATUSIK W., DURAND F.: Multi-aperture photography. *ACM Trans. Graph. 26* (2007). 3

[GZC*06] GEORGEIV T., ZHENG K., CURLESS B., SALESIN D., NAYAR S., INTWALA C.: Spatio-angular resolution tradeoff in integral photography. In *EGSR* (2006), pp. 263–272. 3

[HM98] HIURA S., MATSUYAMA T.: Depth measurement by the multi-focus camera. In *Proc. CVPR* (1998), pp. 953–959. 3

[IWH10] IHRKE I., WETZSTEIN G., HEIDRICH W.: A Theory of Plenoptic Multiplexing. In *Proc. IEEE CVPR* (Jun 2010). oral. 3

[Lev10] LEVIN A.: Analyzing depth from coded aperture sets. In *Proc. ECCV: Part I* (2010), Springer-Verlag, pp. 214–227. 3

[LFDF07] LEVIN A., FERGUS R., DURAND F., FREEMAN W. T.: Image and depth from a conventional camera with a coded aperture. *ACM Trans. Graph. 26* (July 2007). Article 70. 2, 3, 6, 7, 8, 9

[LHG*09] LEVIN A., HASINOFF S. W., GREEN P., DURAND F., FREEMAN W. T.: 4d frequency analysis of computational cameras for depth of field extension. *ACM Trans. Graph. 28* (July 2009), 97:1–97:14. 3

[LLW*08] LIANG C.-K., LIN T.-H., WONG B.-Y., LIU C., CHEN H. H.: Programmable aperture photography: Multiplexed light field acquisition. *ACM Trans. Graph. 27* (2008), 55:1–55:10. 3

[Ng05] NG R.: Fourier slice photography. *ACM Trans. Graph. 24* (July 2005), 735–744. 3

[NKZN08] NAGAHARA H., KUTHIRUMMAL S., ZHOU C., NAYAR S.: Flexible Depth of Field Photography, Oct 2008. 3

[NLB*05] NG R., LEVOY M., BRÉDIF M., DUVAL G., HOROWITZ M., HANRAHAN P.: Stanford tech report ctsr 2005-02 light field photography with a hand-held plenoptic camera, 2005. 3

[NZW*10] NAGAHARA H., ZHOU C., WATANABE T., ISHIGURO H., NAYAR S.: Programmable Aperture Camera Using LCoS. In *Proc. ECCV* (Oct 2010). 3

[RAT06] RASKAR R., AGRAWAL A., TUMBLIN J.: Coded exposure photography: motion deblurring using fluttered shutter. *ACM Trans. Graph. 25*, 3 (2006), 795–804. 1, 2, 8, 9

[VRA*07] VEERARAGHAVAN A., RASKAR R., AGRAWAL A., MOHAN A., TUMBLIN J.: Dappled photography: Mask enhanced cameras for heterodyned light fields and coded aperture refocusing. *ACM Trans. Graph 26* (2007), 3562–3570. Article 69. 2, 3, 6, 8

[WF07] WEISS Y., FREEMAN W.: What makes a good model of natural images. In *Proc. CVPR* (2007), pp. 1–8. 6

[ZLN09] ZHOU C., LIN S., NAYAR S.: Coded aperture pairs for depth from defocus. In *ICCV* (Oct 2009), pp. 325–332. 1, 3, 8

[ZN09] ZHOU C., NAYAR S. K.: What are good apertures for defocus deblurring? In *IEEE Intern. Conf. Computational Photography* (2009), pp. 1–8. 1, 2, 3, 8, 9